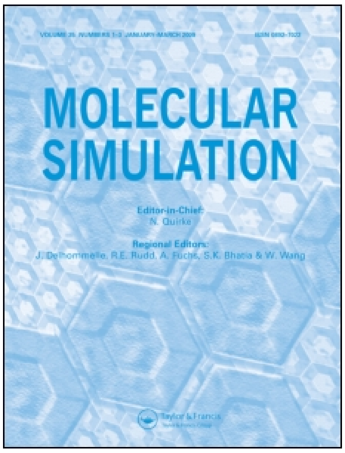

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

On a seven-dimensional representation of RNA secondary structures

B. Liao[a]; T. Wang[b]; K. Ding[c]
[a] Laboratory of Embedded Computing and Systems, School of Computer and Communication, Hunan University, Changsha Hunan, China [b] Department of Applied Mathematics, Dalian University of Technology, Dalian, China [c] Department of Applied Mathematics, Graduate School of the Chinese Academy of Sciences, Beijing, China

To cite this Article Liao, B. , Wang, T. and Ding, K.(2005) 'On a seven-dimensional representation of RNA secondary structures', Molecular Simulation, 31: 14, 1063 — 1071

To link to this Article: DOI: 10.1080/08927020500371332

URL: http://dx.doi.org/10.1080/08927020500371332

Molecular Simulation, Vol. 31, No. 14–15, 15–30 December 2005, 1063–1071

Taylor & Francis
Taylor & Francis Group

On a seven-dimensional representation of RNA secondary structures

B. LIAO†*, T. WANG‡ and K. DING¶

†Laboratory of Embedded Computing and Systems, School of Computer and Communication, Hunan University, Changsha Hunan 410082, China
‡Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, China
¶Department of Applied Mathematics, Graduate School of the Chinese Academy of Sciences, Beijing 100039, China

(Received May 2005; in final form october 2005)

In this paper, we proposed a seven-dimensional (7D) representation of ribonucleic acid (RNA) secondary structures. The use of the 7D representation is illustrated by constructing structure invariants. Comparisons with the similarity/dissimilarity results based on 7D representation for a set of RNA 3 secondary structures at the 3′-terminus of different viruses, are considered to illustrate the use of our structure invariants based on the entries in derived sequence matrices restricted to a selected width of a band along the main diagonal.

Keywords: RNA secondary structure; Similarity; Virus; 7D representation

1. Introduction

In recent years graphical representations of DNA were introduced to facilitate comparison of DNA sequences and servation of differences in their structures. Several authors outlined different approaches to compute the similarity of DNA sequences based on 2D, 3D or 4D graphical representations [2–7,9–11]. The advantage of graphical representation of DNA sequences is that they allow visual inspection of data, helping in recognizing major differences among similar DNA sequences.

Ribonucleic acid (RNA) is an important molecule which performs a wide range of functions in the biological system. In particular, it is RNA (not DNA) that contains genetic information of virus such as HIV and, therefore, regulates the functions of such virus. RNA has recently become the center of much attention because of its catalytic properties, leading to an increased interest in obtaining structural information [8,11–19]. Using the similar methods with the graphical representation of DNA sequences, one also can outline several graphical representations of RNA primary sequences based on 2D, 3D or 4D to compute the similarity of RNA primary sequences. Now, we will consider the similarity of RNA secondary structures.

Previously, almost all such comparisons were based on alignment of RNA structures: a distance function or a score function is used to represent insertion, deletion, and substitution of letters in the compared structures. Using the distance function, one can compute similarity between RNA structures. There are many algorithms for computing the similarity between RNA secondary structures [13–19]. But, by using these approaches, the chemical structures and properties are ignored, and there is a restriction of non-crossing.

In this paper, based on the classifications of bases and base pairs, we shall propose a seven-dimensional (7D) representation, which avoids the limitation associated with non-crossing. We make a comparison of the secondary structures at the 3′-terminus belonging to nine different species based on this graphical representation. In figure 1, the secondary structures at the 3′-terminus belonging to nine different viruses are listed, which were reported by Bol [1]. The similarities are computed by calculating the Euclidean distance between the end points of the vectors or calculating the correlation angle of two vectors.

In Section 2, we introduce the 7D representation of RNA secondary structures. Section 3 presents the structure invariants. In Section 4, we compute the similarities/dissimilarities among nine RNA secondary structures based on bandwidth averages.

*Corresponding author. Fax: + 86-731-8821715. Email: dragonbw@163.com

Molecular Simulation
ISSN 0892-7022 print/ISSN 1029-0435 online © 2005 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/08927020500371332

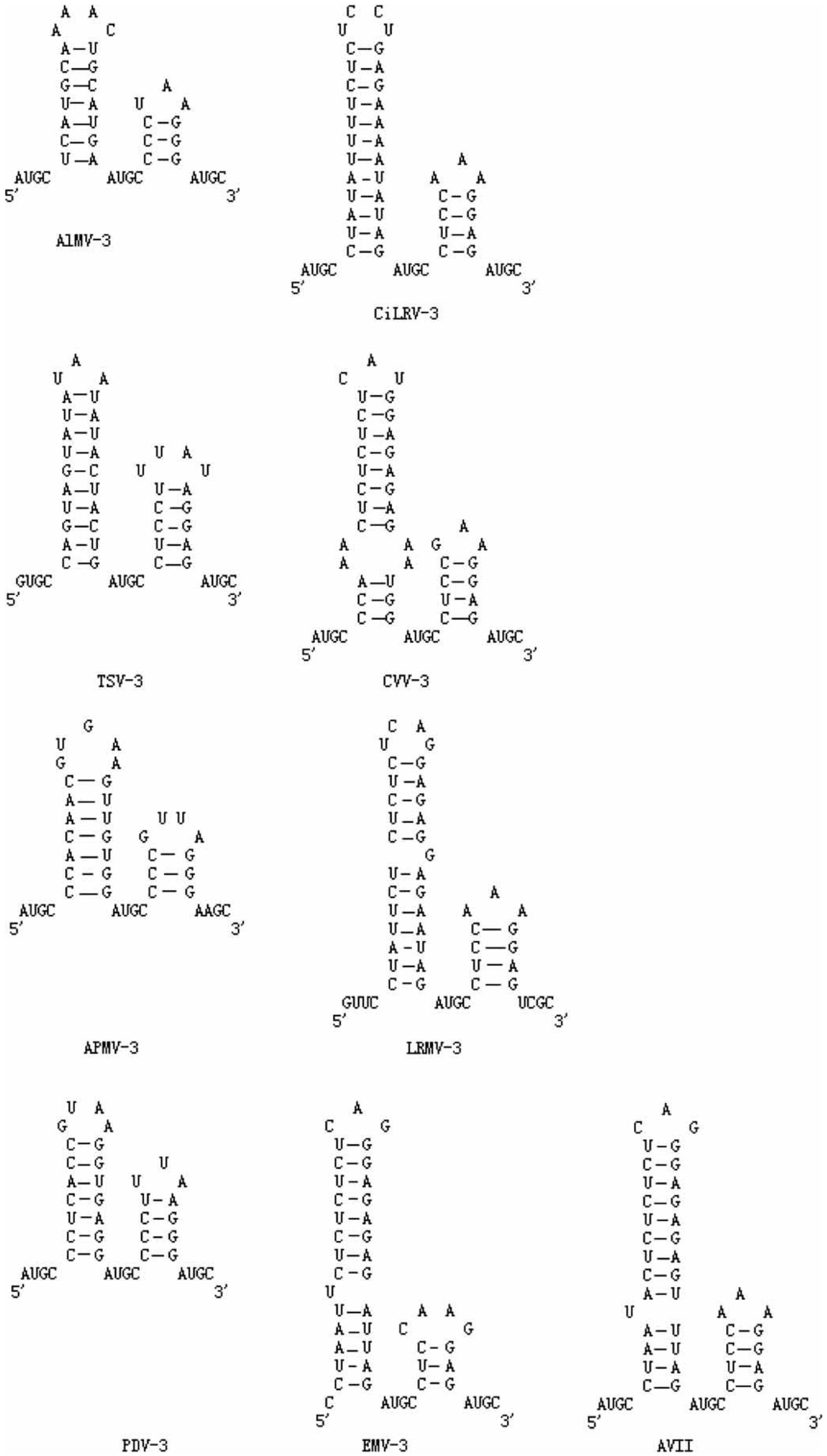

Figure 1. Secondary structure at the 3′-terminus of RNA 3 of alfalfa mosaic virus (AlMV-3 [20]), citrus leaf rugose virus (CiLRV-3 [21]), tobacco streak virus (TSV-3 [22,23]), citrus variegation virus (CVV-3 [21]), apple mosaic virus (APMV-3 [24]), prune dwarf ilarvirus (PDV-3 [25]), lilac ring mottle virus (LRMV-3 [26]), elm mottle virus (EMV-3 [27]) and asparagus virus II (AVII [28]). Numbering of nucleotides is from the 3′end of RNA 3.

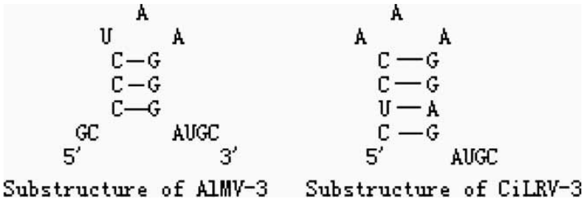

Figure 2. Substructure of AlMV-3 and CiLRV-3.

2. 7D representation of RNA secondary structures

The secondary structure of an RNA is a set of free bases and base pairs forming hydrogen bonds between A–U, G–C, and G–U. Let A', U', G', C', G'', U'' denote A, U, G, C in the base pair A–U, G–C and G–U, respectively. Then we can obtain a special sequence representation of the secondary structure. We call it characteristic sequence of the secondary structure. For example, the corresponding characteristic sequence of the substructure of AlMV-3 (figure 2) is $CGUAG'G'G'AAUC'C'C'CG$ (from 3′ to 5′).

We will illustrate the 7D characterization of RNA secondary structure at the 3′-terminus of RNA 3 listed in figure 1. In 7D space points, vectors and directions have seven components, and we will assign the following basic elementary directions to the four free bases and three base pairs.

$$\begin{array}{rl}
\text{freebase} & \text{A}(1,0,0,0,0,0,0); \\
 & \text{U}(0,1,0,0,0,0,0) \\
 & \text{C}(0,0,1,0,0,0,0); \\
 & \text{G}(0,0,0,1,0,0,0) \\
\text{basepair} & \text{A–U}(0,0,0,0,1,0,0) \\
 & \text{C–G}(0,0,0,0,0,1,0) \\
 & \text{G–U}(0,0,0,0,0,0,1)
\end{array}$$

The above selection is equivalent to any other permutation of labels and directions, because the seven directions of 7D space are fully equivalent, hence, this selection should not be viewed as introducing any arbitrary decision to influence numerical analysis that follows.

We will reduce a RNA secondary structure into a series of nodes $P_0, P_1, P_2, \ldots P_N$, whose coordinates x_n, y_n, z_n, s_n, v_n, w_n, t_n ($n = 0,1,2,\ldots,N$ where N is the length of the RNA secondary structure being studied) satisfy

$$\begin{cases} x_n = A_n \\ y_n = U_n \\ z_n = C_n \\ s_n = G_n \\ v_n = A'_n + U'_n \\ w_n = C'_n + G'_n \\ t_n = G''_n + U''_n \end{cases}$$

Where A_n, C_n, G_n, U_n, A'_n, U'_n, C'_n, G'_n, G''_n and U''_n are the cumulative occurrence numbers of A, C, G, U, A′, U′, C′, G′, G′ and U″, respectively, in the subsequence from the 1st base to the nth base in the sequence. We define $A_0 = C_0 = G_0 = U_0 = A'_0 = U'_0 = C'_0 = G'_0 = G''_n = U''_n = 0$. In table 1, we have introduced 7D coordinates for the first 15 bases of AlMV-3 and CiLRV-3 in order to illustrate differences between the initial stages of the two secondary structures.

Obviously, x_n, y_n, z_n, s_n, v_n, w_n, t_n are independent. So transform $A \Leftrightarrow U$, $A \Leftrightarrow C$, $A \Leftrightarrow G$, $C \Leftrightarrow G$, $C \Leftrightarrow U$, $G \Leftrightarrow U$, $A{-}U \Leftrightarrow C{-}G$, $A{-}U \Leftrightarrow G{-}U$, $G{-}U \Leftrightarrow G{-}C$, $A(U, C \text{ or } G) \Leftrightarrow A{-}U(G{-}C \text{ or } G{-}U)$ means that $x \Leftrightarrow y$, $x \Leftrightarrow z$, $x \Leftrightarrow s$, $z \Leftrightarrow s$, $z \Leftrightarrow y$, $s \Leftrightarrow y$, $v \Leftrightarrow w$, $v \Leftrightarrow t$, $t \Leftrightarrow w$, $x(y, z \text{ or } s) \Leftrightarrow v(w \text{ or } t)$, respectively.

Since we have no graphical representation to associate with a random walk in 7D space, we have constructed the distance matrix for each such random walk in which any (i, j) entry is the Euclidean distance between corresponding points in 7D space given by

$$D_{i,j} = \sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2 + (\Delta s)^2 + (\Delta v)^2 + (\Delta w)^2 + (\Delta t)^2}$$

where $\Delta x = A_i - A_j$, $\Delta y = U_i - U_j$, $\Delta z = C_i - C_j$, $\Delta s = G_i - G_j$, $\Delta v = A'_i + U'_i - A'_j - U'_j$ $\Delta w = C'_i + G'_i - C'_j - G'_j$,

Table 1. 7D Coordinates for the first 15 bases of AlMV-3 and CiLRV-3.

	AlMV-3								*CiLRV-3*							
no:	*Base*	*x*	*y*	*z*	*s*	*v*	*w*	*t*	*Base*	*x*	*y*	*z*	*s*	*v*	*w*	*t*
1	C	0	0	1	0	0	0	0	C	0	0	1	0	0	0	0
2	G	0	0	1	1	0	0	0	G	0	0	1	1	0	0	0
3	U	0	1	1	1	0	0	0	U	0	1	1	1	0	0	0
4	A	1	1	1	1	0	0	0	A	1	1	1	1	0	0	0
5	G'	1	1	1	2	0	1	0	G'	1	1	1	2	0	1	0
6	G'	1	1	1	3	0	2	0	A'	2	1	1	2	1	1	0
7	G'	1	1	1	4	0	3	0	G'	2	1	1	3	1	2	0
8	A	2	1	1	4	0	3	0	G'	2	1	1	4	1	3	0
9	A	3	1	1	4	0	3	0	A	3	1	1	4	1	3	0
10	U	3	2	1	4	0	3	0	A	4	1	1	4	1	3	0
11	C'	3	2	2	4	0	4	0	A	5	1	1	4	1	3	0
12	C'	3	2	3	4	0	5	0	C'	5	1	2	4	1	4	0
13	C'	3	2	4	4	0	6	0	C'	5	1	3	4	1	5	0
14	C	3	2	5	4	0	6	0	U'	5	2	3	4	2	5	0
15	G	3	2	5	5	0	6	0	C'	5	2	4	4	2	6	0

Table 2. A 15 × 15 fragment of the 39 × 39 distance matrix belonging to the 7D representation of AlMV-3.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
1	0	1	$\sqrt{2}$	$\sqrt{3}$	$\sqrt{7}$	$\sqrt{15}$	$\sqrt{27}$	$\sqrt{30}$	$\sqrt{35}$	$\sqrt{38}$	$\sqrt{46}$	$\sqrt{58}$	$\sqrt{74}$	9	$\sqrt{90}$
2		0	1	$\sqrt{2}$	2	$\sqrt{10}$	$\sqrt{20}$	$\sqrt{23}$	$\sqrt{28}$	$\sqrt{31}$	$\sqrt{39}$	$\sqrt{51}$	$\sqrt{70}$	$\sqrt{74}$	9
3			0	1	$\sqrt{3}$	3	$\sqrt{19}$	$\sqrt{22}$	$\sqrt{27}$	$\sqrt{28}$	6	$\sqrt{48}$	8	$\sqrt{71}$	$\sqrt{78}$
4				0	$\sqrt{2}$	$\sqrt{8}$	$\sqrt{18}$	$\sqrt{19}$	$\sqrt{22}$	$\sqrt{23}$	$\sqrt{31}$	$\sqrt{43}$	$\sqrt{59}$	$\sqrt{66}$	$\sqrt{73}$
5					0	$\sqrt{2}$	$\sqrt{8}$	3	$\sqrt{12}$	$\sqrt{13}$	$\sqrt{19}$	$\sqrt{29}$	$\sqrt{43}$	$\sqrt{50}$	$\sqrt{55}$
6						0	$\sqrt{2}$	$\sqrt{3}$	$\sqrt{6}$	$\sqrt{7}$	$\sqrt{11}$	$\sqrt{19}$	$\sqrt{31}$	$\sqrt{38}$	$\sqrt{41}$
7							0	1	2	$\sqrt{5}$	$\sqrt{7}$	$\sqrt{13}$	$\sqrt{23}$	$\sqrt{30}$	$\sqrt{31}$
8								0	1	$\sqrt{2}$	2	$\sqrt{10}$	$\sqrt{20}$	$\sqrt{27}$	$\sqrt{28}$
9									0	1	$\sqrt{3}$	3	$\sqrt{19}$	$\sqrt{26}$	$\sqrt{27}$
10										0	$\sqrt{2}$	$\sqrt{8}$	$\sqrt{18}$	5	$\sqrt{26}$
11											0	$\sqrt{2}$	$\sqrt{8}$	$\sqrt{18}$	$\sqrt{19}$
12												0	$\sqrt{2}$	$\sqrt{5}$	$\sqrt{6}$
13													0	1	$\sqrt{2}$
14														0	1
15															0

$\Delta t = G_i'' + U_i'' - G_j'' - U_j''$. Obviously all transformations do not affect distance measure. So there is a unique distance matrix corresponding to a RNA secondary structure.

In table 2, we show a fragment of 39 × 39 distance matrix corresponding to the 7D representation of AlMV-3. This matrix, which summarizes all the distances between the nucleic bases in the 7D representation, forms our basic information of RNA secondary structures.

We will do mathematical analysis and comparison of distance matrices belonging to different RNA secondary structures. Constructing suitable structural and matrix invariants facilitates such comparisons. In this contribution, we will point out invariants that can be extracted from the distance matrix, such as that shown in table 2.

3. Structure invariants

An invariant, by definition, is a quality (numerical value) that is dependent of any graphical representation of a structure or assignment of labels used to construct the matrix representing structure. Similarity, structure invariants are quantities that can be attributed to a structure independently of labels used to identify individual elements of a structure or any graphical representation of a structure. There is, however, an important difference between structure invariants and matrix invariants in that in a structure the adjacency of neighboring elements is firmly fixed and is not lost in alternative graphical or labelling uses. Thus, for instance, the first two bases C, G of AlMV-3 will remain adjacent regardless of labels used for C and G, which need not be 1 and 2.

Table 3. Expression for band average widths 1–14 for the 15 × 15 fragment of the distance matrix of AlMV-3.

Band	*AlMV-3*
1	$\left(8+6\sqrt{2}\right)/14$
2	$\left(4\sqrt{2}+3\sqrt{3}+4\sqrt{8}+\sqrt{5}+2\right)/13$
3	$\left(13+2\sqrt{6}+3\sqrt{18}+\sqrt{3}+\sqrt{5}\right)/12$
4	$\left(3\sqrt{7}+2\sqrt{10}+4\sqrt{19}+5+\sqrt{12}\right)/11$
5	$\left(\sqrt{15}+2\sqrt{20}+2\sqrt{22}+2\sqrt{13}+2\sqrt{26}+\sqrt{11}\right)/10$
6	$\left(4\sqrt{27}+3\sqrt{23}+2\sqrt{19}\right)/9$
7	$\left(2\sqrt{30}+3\sqrt{28}+\sqrt{29}+2\sqrt{31}\right)/8$
8	$\left(2\sqrt{31}+\sqrt{35}+6+\sqrt{38}+2\sqrt{43}\right)/7$
9	$\left(\sqrt{38}+\sqrt{39}+\sqrt{41}+\sqrt{48}+\sqrt{50}+\sqrt{59}\right)/6$
10	$\left(\sqrt{46}+\sqrt{51}+\sqrt{55}+8+\sqrt{66}\right)/5$
11	$\left(\sqrt{58}+\sqrt{70}+\sqrt{71}+\sqrt{73}\right)/4$
12	$\left(2\sqrt{74}+\sqrt{78}\right)/3$
13	9
14	$\sqrt{90}$

The structure invariant that we will introduce will be explained on the 15 × 15 fragment of the distance matrix of table 2. One can observe that in each row of the table the entries increase from left to right. The matrix can easily be rearranged by first placing the smallest entry (either 1 or $\sqrt{2}$) next to the main diagonal, then the next smallest entry $\left(\sqrt{2}, \sqrt{3}, 2, \sqrt{5}, \text{ or } \sqrt{8}\right)$ next to the first, and so on till all entries are arranged in increasing order as we more from the diagonal zero to the right.

In the following we will assume that all distance matrices considered have already been so ordered. We can now consider the sums of elements in diagonal entries parallel to the main diagonal, which consists of zeroes. In the case of table 2, we obtain for the first few neighboring diagonals the following sums: $8+6\sqrt{2}$, $\left(4\sqrt{2}+3\sqrt{3}+4\sqrt{8}+\sqrt{5}+2\right)$, $\left(13+2\sqrt{6}+3\sqrt{18}+\sqrt{3}+\sqrt{5}\right)$, and so on. We will refer to these sums as band invariants of different width, specifically b_1, b_2, b_3 for $8+6\sqrt{2}$, $\left(4\sqrt{2}+3\sqrt{3}+4\sqrt{8}+\sqrt{5}+2\right)$, $\left(13+2\sqrt{6}+3\sqrt{18}+\sqrt{3}+\sqrt{5}\right)$. One can observe that these quantities are not matrix invariants but can always be extracted from any matrix in whatever form it is presented by considering adjacency between structure elements. If the distance matrix is already in the canonical form based on assigning labels to nucleic acid bases of secondary structures, band invariants can readily be obtained by summing elements along each of the lines parallel to the main diagonal.

In table 3 we show the band invariants based on 15 × 15 fragments of the secondary structure of AlMV-3. One can find that there is considerable variation in the form and numerical values for bandwidth invariants belonging to different secondary structures. In table 4 the numerical values for the average bandwidths are listed.

Table 4. Numerical value for the band average widths 1–14 for the 15 × 15 fragment of the distance matrices for the RNA 3 of nine species of figure 1.

Band	*AlMV-3*	*CiLRV-3*	*TSV-3*	*CVV-3*	*APMV-3*	*LRMV-3*	*PDV-3*	*EMV-3*	*AVII*
0	0	0	0	0	0	0	0	0	0
1	1.1774	1.2366	1.2070	1.2366	1.1774	1.2366	1.2366	1.1774	1.2366
2	2.0308	1.9754	1.8603	1.9303	2.0758	1.9754	2.1352	1.8535	1.9754
3	2.8298	2.7437	2.5691	2.6426	2.8541	2.7095	2.8901	2.5008	2.7437
4	3.5950	3.5008	3.3309	3.3835	3.5096	3.4401	3.5393	3.1944	3.5008
5	4.2924	4.1388	3.9727	4.0276	4.0748	4.0530	4.1764	3.7749	4.1338
6	4.8767	4.8547	4.6896	4.7311	4.6547	4.7616	4.7531	4.3254	4.8547
7	5.4189	5.5604	5.3034	5.4639	5.2080	5.4230	5.2144	4.8598	5.5604
8	6.0471	6.2274	5.8476	6.1593	5.8033	6.0133	5.7546	5.4211	6.2274
9	6.7487	6.9333	6.4022	6.8597	6.5222	6.6538	6.3617	5.9970	6.9333
10	7.4926	7.5706	6.9702	7.4914	7.2916	7.2668	7.0498	6.5224	7.5706
11	8.1928	8.1785	7.5307	8.1167	7.9707	7.9230	7.7340	7.1175	8.1785
12	8.6787	8.6920	8.0803	8.6157	8.5017	8.4210	8.2590	7.5900	8.6920
13	9	9.0795	8.6890	9.0260	9.0550	8.9110	8.8560	8.0620	9.0795
14	9.4870	9.6950	9.1100	9.6950	9.4870	9.5920	9.4870	8.4850	9.6950

4. Similarities/dissimilarities based on bandwidth averages

We will illustrate the use of the 7D quantitative characterization of RNA secondary structure with the examination of similarities/dissimilarities among the nine structures listed in figure 1. We construct a n-component vectors (with $n = 5$, 10, or 15) consisting of the average bandwidth. The underlying assumption is that if two vectors point to a similar direction in the n-dimensional space, and then the two RNA secondary structure represented by the n-component vectors are similar.

The similarities among such vectors can be computed in three ways: (1) we calculate the Euclidean distance between the end points of the vectors; (2) we calculate the correlation angle of two vectors, and (3) we calculate the cosine of the correlation angle of two vectors. When one calculates the correlation angle of two vectors, the cosine of the correlation angle of two vectors can be obtained easily. The smaller Euclidean distance between the end points of two vectors, the more similar the RNA secondary structure. And, the smaller correlation angle between two vectors, the more similar the RNA secondary structure. On the other hand, the larger the cosine of the correlation angle between two vectors, the more similar the RNA secondary structure.

In tables 5 – 10 we have listed the similarity/dissimilarity table between the secondary structure at the $3'$-terminus of RNA 3 of nine species based on cumulative bandwidths of order 5, 10, and 15, that is based on vectors with 5, 10, and 15 components. The three different bandwidths that we consider correspond to invariants that consider RNA substructure with length 5, 10, and 15, all confined to the first 15 RNA nucleic acid bases. Observing tables 5–10, we can find that the smallest entry 0 indicates that the substructures of CiLRV-3 with length 5, 10, or 15 are the same as the corresponding substructures of AVII. Figure 1 proved the results.

A comparison of the three separate similarities may vary. Such as, for the shortest segments, in table 6 the most similar are CVV-3-EMV-3 (except for CilRV-3-AVII) with the smallest value 0.006029, but as we increase the length of the segments, the similarity of the pair CVV-3-EMV-3 decreases considerably. On the other hand, the similarity between the pairs CiLRV-3-AVII, LRMV-3-CVV-3, CVV-3-AVII and LRMV-3-AVII is only slightly affected by the increase of the length of the fragment considered and corresponds to the smaller numerical entry for the cumulative bandwidths based on all 15 components of the vectors in table 4.

In tables 11–13 we show the similarity/dissimilarity table for the nine secondary structure based on cumulative bandwidths of order 5 but using the full structure listed in figure 1. Interestingly, with the full RNA secondary

Table 5. Similarity/dissimilarity between the nine species of figure 1 based on the Euclidean distances between the end points of the 5-component vectors of the bandwidth 5.

Species	*AlMV-3*	*CiLRV-3*	*TSV-3*	*CVV-3*	*APMV-3*	*LRMV-3*	*PDV-3*	*EMV-3*	*AVII*
AlMV-3	0	0.151197	0.409463	0.305583	0.099542	0.212226	0.145405	0.547866	0.151179
CiLRV-3		0	0.271063	0.161290	0.160781	0.069672	0.220116	0.413818	0
TSV-3			0	0.118090	0.400593	0.213918	0.472142	0.155626	0.271063
CVV-3				0	0.292076	0.098555	0.357091	0.255478	0.161290
APMV-3					0	0.198303	0.095974	0.523058	0.160781
LRMV-3						0	0.260755	0.349698	0.069672
PDV-3							0	0.594449	0.220116
EMV-3								0	0.413818
AVII									0

Table 6. The similarity/dissimilarity matrix for the nine species of figure 1 based on the angle between the end points of the 5-component vectors of the bandwidth 5.

Species	*AlMV-3*	*CiLRV-3*	*TSV-3*	*CVV-3*	*APMV-3*	*LRMV-3*	*PDV-3*	*EMV-3*	*AVII*
AlMV-3	0	0.017905	0.027100	0.026510	0.018645	0.021729	0.026363	0.029127	0.017905
CiLRV-3		0	0.010507	0.009334	0.025869	0.006784	0.027000	0.013837	0
TSV-3			0	0.008021	0.035945	0.012400	0.035601	0.013469	0.010507
CVV-3				0	0.031043	0.006244	0.028927	0.006029	0.009334
APMV-3					0	0.024809	0.010713	0.030590	0.025869
LRMV-3						0	0.023268	0.008120	0.006784
PDV-3							0	0.026603	0.027000
EMV-3								0	0.013837
AVII									0

structure, the most similar are EMV-3 and AVII with the lowest value of 0.001730, followed by AVII and LRMV-3 with a value of 0.004318 and by LRMV-3 and EMV-3 with a value of 0.005735.

The Euclidean distance between the end points of vectors and the correlation angle between vectors are different measures of the similarity of RNA secondary structures. There exists an overall qualitative agreement among nine species' similarities. However, there are small differences between tables 5 and 6, 7 and 8, 9 and 10, and 12 and 13. The reasons for these differences may be as follows: (1) there is some loss of information associated with the distance matrix and obtained by calculating the average values for matrix. (2) information extracted in each structure is not plenteous enough to allow comparison of nine species. In general, the correlation angle is the best tolerance for the similarities.

The similarities based on the full RNA secondary structure are somewhat altered when the full RNA secondary structure is compared with the initial part of the structure, as shown in figure 3. Entries that remain close to the line $y = x$ indicate pairs of RNA structures which have not varied much after the initial changes. Such are, for instance, the pairs {AVII, LRMV-3}, {PDV-3, APMV-3}, {AVII, CVV-3}. Entries that are at greater distance from the line $y = x$ indicate RNA secondary structures that show considerable variation not only at the initial fragments, but also throughout their length. In figure 3, we show a plot of bandwidths 1–5 for the initial fragment of length 15 of AlMV-3 against the bandwidths 1–5 for the full length of AlMV-3.

5. Conclusions

High complexity and degeneracy are major problems in previous RNA secondary structure representations. Our representation provides a direct plotting method to denote RNA secondary structures. From the RNA representation, the A, U, G, C, A-U, G-u and C-G usage as well as the original RNA structure can be recaptured mathematically without loss of textual information. Many properties of visual importance in a RNA secondary structure are preserved in the 7D representation. It is useful for visualizing the local and global features of big or small RNA secondary structures and can facilitate the visual discovery of interesting features in a RNA secondary structure. Another advantage of our approach is that it allows visual inspection of data, helping in recognizing major similarities among different RNA structures and it avoids the limitation of non-crossing. It is well-known that the alignments of RNA secondary structures are computer intensive that is direct comparison for RNA secondary structure. Structure considered in alignment of RNA structures is only string's structures. Here, we use an intensive approach which shall consider not only sequences' structure but also chemical structure for RNA secondary structures. The structure invariant easily computed and compared is applied to compare RNA

Table 7. Similarity/dissimilarity between the nine species of figure 1 based on the Euclidean distances between the end points of the 10-component vectors of the bandwidth 10.

Species	*AlMV-3*	*CiLRV-3*	*TSV-3*	*CVV-3*	*APMV-3*	*LRMV-3*	*PDV-3*	*EMV-3*	*AVII*
AlMV-3	0	0.365438	0.691431	0.460037	0.511627	0.354634	0.571858	1.463358	0.365438
CiLRV-3		0	0.787837	0.270229	0.736883	0.404625	0.854500	1.612900	0
TSV-3			0	0.592372	0.444501	0.403059	0.535401	0.859567	0.787837
CVV-3				0	0.632060	0.276897	0.790048	1.395653	0.270229
APMV-3					0	0.398586	0.239560	1.007533	0.736883
LRMV-3						0	0.528256	1.220320	0.404625
PDV-3							0	1.033184	0.854500
EMV-3								0	1.612900
AVII									0

Table 8. The similarity/dissimilarity matrix for the nine species of figure 1 based on the angle between the end points of the 10-component vectors of the bandwidth 10.

Species	*AlMV-3*	*CiLRV-3*	*TSV-3*	*CVV-3*	*APMV-3*	*LRMV-3*	*PDV-3*	*EMV-3*	*AVII*
AlMV-3	0	0.025272	0.022288	0.033391	0.015590	0.019115	0.024907	0.012961	0.025272
CiLRV-3		0	0.016161	0.008709	0.032251	0.010575	0.045924	0.023061	0
TSV-3			0	0.022438	0.032270	0.010497	0.039727	0.018830	0.016161
CVV-3				0	0.039741	0.018235	0.053708	0.030276	0.008709
APMV-3					0	0.025384	0.018495	0.016799	0.032251
LRMV-3						0	0.036738	0.014433	0.010575
PDV-3							0	0.024807	0.045924
EMV-3								0	0.023061
AVII									0

Table 9. Similarity/dissimilarity between the nine species of figure 1 based on the Euclidean distances between the end points of the 15-component vectors of the bandwidth 15.

Species	*AlMV-3*	*CiLRV-3*	*TSV-3*	*CVV-3*	*APMV-3*	*LRMV-3*	*PDV-3*	*EMV-3*	*AVII*
AlMV-3	0	0.435425	1.336521	0.515108	0.621164	0.578704	0.964611	2.703348	0.435425
CiLRV-3		0	1.5065541	0.302984	0.862610	0.658588	1.216424	2.922276	0
TSV-3			0	1.307057	0.973726	0.895551	0.732306	1.460912	1.506554
CVV-3				0	0.719609	0.475793	1.079372	2.707520	0.302984
APMV-3					0	0.447252	0.519988	2.270681	0.862610
LRMV-3						0	0.634160	2.308452	0.658588
PDV-3							0	1.951331	1.216424
EMV-3								0	2.922276
AVII									0

Table 10. The similarity/dissimilarity matrix for the nine species of figure 1 based on the angle between the end points of the 15-component vectors of the bandwidth 15.

Species	*AlMV-3*	*CiLRV-3*	*TSV-3*	*CVV-3*	*APMV-3*	*LRMV-3*	*PDV-3*	*EMV-3*	*AVII*
AlMV-3	0	0.016062	0.021035	0.021965	0.018375	0.018288	0.024869	0.014194	0.016062
CiLRV-3		0	0.015698	0.008159	0.023449	0.010821	0.030440	0.015986	0
TSV-3			0	0.019551	0.024828	0.011806	0.024972	0.012069	0.015698
CVV-3				0	0.024773	0.012663	0.033388	0.021311	0.008159
APMV-3					0	0.019257	0.017804	0.018684	0.023449
LRMV-3						0	0.022105	0.012365	0.010821
PDV-3							0	0.018544	0.030440
EMV-3								0	0.015986
AVII									0

Table 11. Initial band average width for the distance matrices of the nine species of figure 1.

Band	*AlMV-3*	*CiLRV-3*	*TSV-3*	*CVV-3*	*APMV-3*	*LRMV-3*	*PDV-3*	*EMV-3*	*AVII*
1	1.2179	1.2650	1.2588	1.2435	1.2070	1.2598	1.2277	1.2760	1.2707
2	1.9626	2.0657	1.9611	1.9122	1.9604	1.9717	2.0013	1.9781	1.9714
3	2.6163	2.9172	2.7002	2.6467	2.6691	2.7642	2.7392	2.7726	2.7728
4	3.2867	3.7438	3.5057	3.3181	3.3709	3.5230	3.4444	3.4950	3.5016
5	3.9756	4.5567	4.2516	4.0516	4.0229	4.2743	4.1271	4.2430	4.2485

Table 12. Similarity/dissimilarity between the nine species of figure 1 based on the Euclidean distances between the end points of the 5-component vectors of the average bandwidth 5 using the full RNA secondary structure.

Species	*AlMV-3*	*CiLRV-3*	*TSV-3*	*CVV-3*	*APMV-3*	*LRMV-3*	*PDV-3*	*EMV-3*	*AVII*
AlMV-3	0	0.806230	0.364488	0.104315	0.110627	0.410819	0.254007	0.378070	0.384726
CiLRV-3		0	0.455902	0.730439	0.707109	0.400967	0.558049	0.434756	0.428281
TSV-3			0	0.284045	0.272260	0.070880	0.152847	0.077557	0.074464
CVV-3				0	0.086116	0.330441	0.195951	0.298626	0.304282
APMV-3					0	0.313522	0.152562	0.276765	0.282490
LRMV-3						0	0.174291	0.046235	0.036283
PDV-3							0	0.141350	0.147925
EMV-3								0	0.012117
AVII									0

Table 13. The similarity/dissimilarity matrix for the nine species of figure 1 based on the angle between the end points of the 5-component vectors of the average bandwidth 5 using the full RNA secondary structure.

Species	*AlMV-3*	*CiLRV-3*	*TSV-3*	*CVV-3*	*APMV-3*	*LRMV-3*	*PDV-3*	*EMV-3*	*AVII*
AlMV-3	0	0.028949	0.022049	0.012703	0.009230	0.020154	0.009968	0.016279	0.017763
CiLRV-3		0	0.013151	0.021960	0.023912	0.011459	0.022422	0.016625	0.015145
TSV-3			0	0.014356	0.019404	0.007003	0.018937	0.011038	0.010264
CVV-3				0	0.012717	0.011208	0.012660	0.006237	0.007887
APMV-3					0	0.016127	0.002471	0.012522	0.013456
LRMV-3						0	0.015020	0.005735	0.004318
PDV-3							0	0.012050	0.012912
EMV-3								0	0.001730
AVII									0

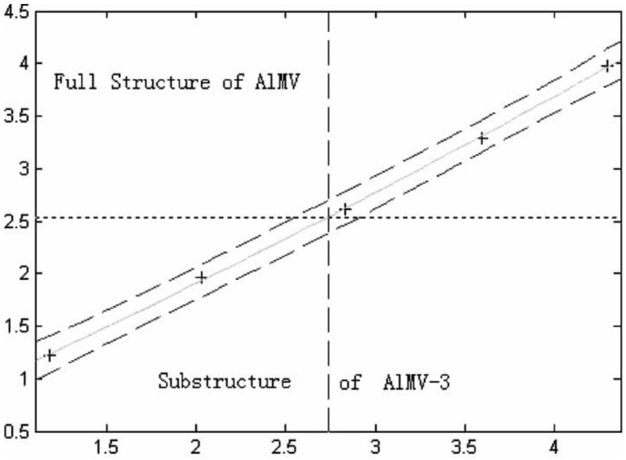

Figure 3. Plot of initial average bandwidths 1–5 as computed from the complete RNA secondary structure of AlMV-3 versus the corresponding bandwidths of the substructure with length 15.

secondary structures, rather than strings' structure themselves. The current 7D graphical representation of RNA secondary structures provides different approaches for both computational scientists and molecular biologists to analysis RNA secondary structures efficiently.

Acknowledgements

The authors would like to thank the anonymous referees for many valuable suggestions that have improved this manuscript.

References

[1] C.B.E.M. Reusken, J.F. Bol. Structural elements of the 3′-terminal coat protein binding site in alfalfa mosaic virus RNAs. *Nucl. Acids Res.*, **14**, 2660 (1996).

[2] M. Randic, M. Vracko, N. Lers, Dejanplavsic. Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation. *Chem. Phys. Lett.*, **371**, 202 (2003).

[3] M. Randic, M. Vracko. On the similarity of DNA primary sequences. *J. Chem. Inf. Comput. Sci.*, **40**, 599 (2000).

[4] M. Randic. Condensed representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.*, **40**, 50 (2000).

[5] M. Randic, A.T. Balanba. On a four-dimensional representation of DNA primary sequences. *J. Chem. Inf. Comput. Sci.*, **40**, 50 (2000).

[6] C. Yuan, B. Liao, T. Wang. New 3-D graphical representation of DNA sequences and their numerical characterization. *Chem. Phys. Lett.*, **379**, 412 (2003).

[7] B. Liao, T. Wang. Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation. *Chem. Phys. Lett.*, **388**, 195 (2004).

[8] B. Liao, T. Wang. General combinatorics of RNA hairpins and cloverleaves. *J. Chem. Inf. Comput. Sci.*, **43**(4), 1138 (2003).

[9] B. Liao, T. Wang. 3-D graphical representation of DNA sequences and their numerical characterization. *J. Mol. Struct. (Theochem)*, **681**, 209 (2004).

[10] B. Liao, T. Wang. New 2D Graphical representation of DNA sequences. *J. Comp. Chem.*, **25**(11), 1364.

[11] B. Liao, T. Wang. Analysis of similarity of DNA sequences based on triplets. *J. Chem. Inf. Comput. Sci.*, **44**, 1666 (2004).

[12] B. Liao, T. Wang. General combinatorics of RNA secondary structure. *Math. Biosci.*, **191**, 69 (2004).

[13] V. Bafna, S. Muthukrisnan, R. Ravi. Comparing similarity between RNA strings. *Comput. Sci.*, **937**, 1 (1995).

[14] F. Corpet, B. Michot. RNA lign program: Alignment of RNA sequences using both primary and secondary structures. *Computer. Appl. Biosci.*, **10**(4), 389 (1995).

[15] S.Y. Le, R. Nussinov, J.V. Mazel. Tree graphs of RNA secondary structures and their comparison. *Comput. Biomed. Res.*, **22**, 461 (1989).

[16] S.Y. Le, J. Onens, R. Nussinov, J.H. Chen, B. Shapiro, J.R. Mazel. RNA secondary structures: Comparison and determination of frequently recurring sunstructures by consensus. *Computer Biomed.*, **5**, 205 (1989).

[17] B. Shapiro. An algorithm for comparing multiple RNA secondary structures. *Comput. Appl. Biosci.*, **4**(3), 387 (1998).

[18] B. Shapiro, K. Zhang. Comparing multiple RNA secondary structures using tree comparisons. *Computer. Appl. Biosci.*, **6**(4), 309 (1990).

[19] K. Zhang. Computing similarity between RNA secondary structures. *Proceeding of the IEEE International Joint Symposia on Intelligence and Sytems*, Rockviue, Maryland 126 (1998).

[20] E.C. Koper-Zwarthoff, F.Th. Brederode, P. Walstra, J.F. Bol. *Nucl. Acids Res.*, **7**, 1887 (1979).

[21] S.W. Scott, X. Ge. The complete nucleotide sequence of RNA 3 of citrus leaf rugose and citrus variegation ilariruses. *J. Gen. Virol.*, **76**, 957 (1995).

[22] E.C. Koper-Zwarthoff, F.Th. Brederode, P. Walstra, J.F. Bol. Nucleotide sequence of the putative recognition site for coat protein in the RNAs of alfalfa mosaic virus and tobacco streak virus. *Nucl. Acids Res.*, **8**, 3307 (1980).

[23] B.J.C. Cornelissen, H. Janssen, D. Zuidema, J.F. Bol. Complete nucleotide sequence of tobacco streak virus RNA. *Nucl. Acids Res.*, **12**, 2427 (1984).

[24] R.H. Alrefai, P.J. Shicl, L.L. Domier, C.J. D'Arcy, P.H. Berger, S.S. Korban. The nucleotide sequence of apple mosaic virus coat protein gene has no similarity with other Bromoviradae coat protein genes. *J. Gen. Virol.*, **75**, 2847 (1994).

[25] S.W. Scott, X. Ge. The complete nucleotide sequence of the RNA 3 of lilac ring mottle ilarvirus. *J. Gen. Virol.*, **76**, 1801 (1995).

[26] E.J. Bachman, S.W. Scott, G. Xin, V. Bowman Vance. The Complete Nucleotide Sequence of Prune Dwarf Ilarvirus RNA 3: Implications for Coat Protein Activation of Genome Replication in Ilarviruses. *Virology*, **201**, 127 (1994).

[27] F. Houser-Scott, M.L. Baer, K.F. Liem, J.M. Cai, L. Gehrke. Nucleotide sequence and structural determinants of specific binding of coat protein or coat protein peptides to the 3 untranslated region of alfalfa mosaic virus RNA 4. *J. Virol.*, **68**, 2194 (1994).

[28] EMBL/GenBank/DDBJ databases. Accession no. X86352.